Chinese-Catalan: A Neural Machine Translation Approach based on Pivoting and Attention Mechanisms

MARTA R. COSTA-JUSSÀ, NOÉ CASAS, CARLOS ESCOLANO AND JOSÉ A. R. FONOLLOSA, Universitat Politècnica de Catalunya, Spain

This paper innovatively addresses machine translation from Chinese to Catalan using neural pivot strategies trained without any direct parallel data. The Catalan language is very similar to Spanish from a linguistic point of view, which motivates the use of Spanish as pivot language. Regarding neural architecture we are using the latest state-of-the-art which is the Transformer model, only based on attention mechanisms. Additionally, this work provides new resources to the community which consist on a human developed gold standard of 4,000 sentences between Catalan and Chinese and all the others United Nations official languages (Arabic, English, French, Russian and Spanish). Results show that the standard pseudo-corpus or synthetic pivot approach performs better than cascade.

CCS Concepts: • **Computing methodologies → Machine translation**;

Additional Key Words and Phrases: Neural Machine Translation, Pivot Approaches, Chinese-Catalan, Transformer

ACM Reference Format:
Marta R. Costa-jussà, Noé Casas, Carlos Escolano and José A. R. Fonollosa. 2017. Chinese-Catalan: A Neural Machine Translation Approach based on Pivoting and Attention Mechanisms. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 9, 4, Article 39 (March 2017), 8 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

In machine translation, low-resource language pairs are those where the available parallel corpora are scarce or not large enough. In some of these cases, despite the absence of directly translated corpora, there is availability of parallel corpora of each of the languages in the pair with a third language, that is, for languages A and B with low availability of A-B parallel corpora, there is a third language P (pivot) for which there are parallel corpora for pairs A-P and B-P. In this situation, it is possible to use such parallel corpora to devise machine translation systems for language pair A-B. The techniques that make it possible are referred to as *pivotal machine translation* techniques, as they use language P as pivot to make the translation between A and B possible. Although these techniques have been widely explored for statistical machine translation [13], and have also been recently explored on the basic recurrent neural machine translation architecture [6], there have not been yet experimented for the case of latest neural machine translation architectures such as the Transformer model [28].

This paper brings together the standard pivotal machine translation techniques (cascade and pseudo corpus) for the latest architecture of neural machine translation, the Transformer, for the

Author's address: Marta R. Costa-jussà, Noé Casas, Carlos Escolano and José A. R. Fonollosa, Universitat Politècnica de Catalunya, C/Jordi Girona, Barcelona, 08034, Spain, \protect\T1\textbraceleftmarta.ruiz,noe.casas,carlos.escolano,jose.fonollosa\protect\T1\textbraceright@upc.edu.

2375-4699/2017/3-ART39 $15.00
https://doi.org/0000001.0000001

specific case of Chinese-Catalan. This becomes the first work on both this pair of languages and, reporting a comparison using the two standard pivot techniques with the Transformer.

Additionally, together with the previous novelties, we are releasing the first Catalan translation gold standard with all the official United Nations languages (Arabic, Chinese, French, Russian and Spanish). This gold standard, which contains 4,000 sentences, is the same as the one provided for other languages in the release of the *United Nations v1.0* [29].

The paper is organised as follows. Section 2 reports linguistically interesting properties of both Catalan as a language and Chinese-to-Catalan translation task. Section 3 reports the most relevant state-of-the-art in Chinese-Catalan and neural machine translation pivot approaches. Section 4 reports a brief description of the neural machine translation architectures. Section 5 describes the classical pivotal machine translation techniques that we apply to neural machine translation. Section 6 reports the experimental details and results and finally, section 7 outlines the conclusions of this work.

2 MOTIVATION OF THE TASK

There is substantial economic interest behind Catalan and Chinese cultures. Catalonia has currently 7.5 million population out of which 20,000 are Chinese[1], which accounts for the second most relevant foreigner community in Catalonia. Given the large distance between both languages, automatic translation entails a great asset in the context of the Catalan society. Additionally, the commercial relationship between these two communities is only growing [2]. Other examples includes the fact that Catalonia attracted 40% of Chinese investment received by Spain in 2015 [5] or that in the period 2005-2050 China will be one of the countries that will generate more immigration, with the US and Spain being the two main recipients [1]. Beyond economical reasons, translation between these two languages is relevant for political[3] and scientific reasons (as we are arguing in the following paragraph).

From the scientific and linguistic point of view, challenges between Chinese and Catalan translation are highly interesting since current state-of-the-art is not yet offering solutions. At the morphological level, Chinese is an analytical language meaning that it has a low morpheme-per-word ratio [15]. On the other hand, Catalan is a highly inflected language having at least one independent morpheme per word and these morphemes are mixed together in words with no clear limits. At the lexical level, Chinese is a language with a massive number of homonyms, which added to the lack of morphological inflections makes the lexical semantic disambiguation towards Catalan even harder. At the syntax level, Chinese and Catalan follow the Subject-Verb-Object pattern and, theoretically, this narrows the reordering. At the end of the day, challenges are very similar to Chinese and Spanish [11]. The linguistic similarity between Spanish and Catalan [12] together with the availability of a large corpus from Spanish-Catalan are the main reasons why we are using Spanish as pivot language in this work.

3 RELATED WORK

In this section we report related work regarding the Chinese-Catalan task and the pivot approach.

Chinese-Catalan. There are no previous works in the language pair of Chinese-Catalan. However, we can find research for Chinese-Spanish and Spanish-Catalan language pairs. Without aiming at

[1]Data taken from https://www.idescat.cat/
[2]https://www.elperiodico.cat/ca/economia/20160102/asia-xina-activitat-importacio-exportacio-port-barcelona-contenidors-4784599
[3]https://www.forbes.com/sites/davidhutt/2017/11/09/why-the-catalonia-independence-crisis-matters-in-beijing/#266e86203494

completeness, works for both language pairs include rule-based [10, 24], statistical-based [11, 16] and neural-based approaches [9, 10]. Pivot approaches (for Chinese-Spanish) have only been studied in the case of statistical-based approaches [13].

Neural machine translation with pivot approaches. Neural machine translation is capable of outperforming statistical machine translation when having a large quantity of data available [20]. However, recently there have been many approaches that focus on tackling neural machine translation with little or non-parallel corpus by training only on monolingual data [2, 21] or on other language parallel corpora [19].

Few works make use of pivot approaches in machine translation. Cheng et al. [6] propose a model to joint training the source-to-pivot and pivot-to-target models. With this approach authors improve over 9% relative BLEU on two language pairs from the WMT benchmark database[4]. Costa-jussà et al. [14] use the cascade pivot strategy to translate from English to Catalan in the biomedical domain.

4 NEURAL MACHINE TRANSLATION WITH ATTENTION MECHANISMS

Early neural machine translation models where designed based on the encoder-decoder architecture [26], usually referred to as sequence-to-sequence model, where the encoder part consisted in a recurrent unit (normally a Long-Short Term Memory [18] or a Gated Recurrent Unit [7]) that receives the embedded input sequence tokens and condensed it into a hidden fixed-size representation that is received by the encoder as initial thought vector, together with the embedded target tokens at the positions before each token prediction.

In sequence-to-sequence models, the information from the source sentence is therefore passed through a fixed-size *bottleneck* representation received by the decoder, independently from the actual length of the source sentence. The introduction of attention mechanisms in [3, 22] allowed this type of models to overcome such a bottleneck, by making the decoding of each target token focus dynamically on specific tokens of the source sentence. This enabled such kind of neural machine translation models to process long sentences and surpass the translation quality of statistical machine translation in several language pairs [4, 27], especially for morphologically-rich languages and pairs with considerable reordering.

The current state-of-the-art neural machine translation architecture is the Transformer model [28], which makes use of a static form of attention based on the dot product of the internal representations, together with linear projections of inputs and outputs and a residual connection. This dot product attention is replicated and assembled into a multi-head attention block. If the block is conditioned on the same sequence as the one used as input, it is referred to as *self-attention*, and its purpose is to draw the dependencies among such input sequence tokens. The Transformer architecture consists of an encoder-decoder setup where both encoder and decoder comprise several multi-head attention layers, and a cross encoder-decoder connection with a multi-head attention block receiving as input the output of the encoder and conditioning on the result of previous layers of the encoder. The self-attention blocks in the decoder are slightly modified to avoid the model having access to tokens that appear at the position of or after the currently predicted token. In order to capture positional information of the words, the Transformer model also incorporates a variation of positional embeddings [17] that uses a sinusoid to also reflect information locality.

5 PIVOT ALTERNATIVES AND NEURAL MACHINE TRANSLATION ARCHITECTURE

Pivot translation strategies are used for low resource language pair A-B translation, when there is a third language P (i.e. the pivot language) for which there is parallel training data with both A and

[4]http://www.statmt.org/wmt18/

B languages. There are different strategies to profit from the parallel data of A and B with P. In this section, we describe the classical pivot strategies which have been used in statistical machine translation and which are studied in this work, for which language A is Chinese, language B is Catalan and the pivot language P is Spanish.

The **cascade approach** consists in training two different translation systems from A to P (pivot language) and from P to B. Then, in inference, two translations have to be performed. This approach is the baseline system in work by Cheng et al. [6]

The **pseudo-corpus or synthetic approach** consists in training a translation system from P to B and using it to translate the entire P side from corpus A-P into B, therefore obtaining an A-B pseudo-corpus where the B side is synthetic. Then, the task is to train the A-B translation system using the pseudo-corpus as training data. An alternative formulation is to train a translation system from P to A and then use this system to translate the entire P side from corpus P-B into A, obtaining an A-B pseudo-corpus where the A side is synthetic. With the pseudo-corpus approach only one translation has to be performed at inference time.

In translation of language pairs that are similar (Chinese-Spanish) to the one studied in this work (Chinese-Catalan), the pseudo-corpus approach achieves better results than the cascade approach for statistical machine translation [13].

6 EXPERIMENTS

In this section we report the details on the experiments proposed. Subsections includes the data and systems details and results.

6.1 Data and System details

Resources come from two main sources: Chinese-Spanish database from the *United Nations v1.0* release [29]; and the Spanish-Catalan corpus is extracted from ten years of the paper edition of a bilingual Catalan newspaper, *El Periódico* [12]. This Spanish-Catalan corpus is partially available via ELDA (Evaluations and Language Resources Distribution Agency). Since we required a gold standard in Chinese-Catalan, we translated the Spanish test set from the *United Nations v1.0* [29] into Catalan. The translation was performed in two steps: we did a first automatic translation from Spanish to Catalan and then a professional translator postedited the output. This Chinese-Catalan test set is freely available upon request to authors. The size of the corpora is summarized in Table 1 for the training data and in Table 2 for the test data. Note that Zh-Ca is the new data set provided in this work and available under request.

Table 1. Size of the parallel training corpora

Language Pair	**Corpus**	Language	Segments	Words	Vocab
Zh-Es	United Nations	Zh	$15.4 \cdot 10^6$	$380.4 \cdot 10^6$	$613 \cdot 10^3$
		Es		$493.4 \cdot 10^6$	$817 \cdot 10^3$
Es-Ca	El Periódico	Es	$6.5 \cdot 10^6$	$165.1 \cdot 10^6$	$736 \cdot 10^3$
		Ca		$178.9 \cdot 10^6$	$713 \cdot 10^3$

The corpora has been pre-processed with a standard pipeline for Catalan and Spanish: tokenizing and keeping parallel sentences between 1 and 50 words. Additionally, for English and Spanish we used Freeling [23] to tokenize pronouns from verbs (i.e. *preguntándose* to *preguntando* + *se*), we also split prepositions and articles, i.e. *del* to *de* + *el* and *al* to *a* + *el*. For Spanish and Catalan, we used Freeling to tokenize the text but no split with pronouns, prepositions or articles was done.

Table 2. Size of the parallel test sets

Language Pair	**Corpus**	Language	Segments	Words	Vocab
Zh-Es	United Nations	Zh	4000	$103.8 \cdot 10^3$	$9.3 \cdot 10^3$
		Es		$139.0 \cdot 10^3$	$12.0 \cdot 10^3$
Es-Ca	El Periódico	Es	2244	$56.0 \cdot 10^3$	$12.2 \cdot 10^3$
		Ca		$60.7 \cdot 10^3$	$11.7 \cdot 10^3$
Zh-Ca	United Nations	Zh	4000	$103.8 \cdot 10^3$	$9.3 \cdot 10^3$
		Ca		$12.3 \cdot 10^3$	$15.7 \cdot 10^3$

For the Chinese corpus, a previous segmentation step is needed in order to identify word boundaries, which are not normally reflected in written Chinese. In order to perform such a processing, we rely on Jieba[5].

The input and expected output of the neural machine translation model are tokens from a closed a priori defined vocabulary. For the three languages, Chinese, Spanish and Catalan, such vocabulary was prepared by means of the variation of Byte-Pair Encoding (BPE) [25] used in [28], which takes the words in the training data and, starting with every possible character as vocabulary, iteratively joins those tokens that most frequently appear together, until a specific target vocabulary size is reached. For morphologically-rich languages, BPE aims at statistically capture the morphological variations of the words while, in the case of Chinese, BPE helps addressing the abundance of multi-character words whose combined semantics can be derived from their parts.

For the Spanish-to-Catalan translation systems, the BPE vocabulary is shared by both source and target sides, meaning that the data used to to extract the vocabulary is the combination of the training data of both sides. Furthermore, the embedding space is also shared between input and output, which serves as regularization due to the reduction in the number of parameters.

The Transformer model used in this work is the original authors' implementation, which is provided as part of `tensor2tensor` [6]. A standard `tensor2tensor` configuration was used, consisting in exponentially-decaying dynamic learning rate and the Adam optimizer. The complete hyperparameter list used for all the attention-based neural machine translation models in this work is shown in Table 3.

Table 3. Hyperparameters of the neural model.

Hyperparameter	Value
number of multi-head attention layers	6
number of attention heads per layer	8
hidden size (embedding)	512
batch size (in tokens)	4096 (× 4 GPU)
training steps	250000
tokenization strategy	BPE
target vocabulary size	32K

6.2 Results

Table 4 shows the BLEU results for all the systems involved in the Chinese-to-Catalan translation, where *pseudo-corpus1* is prepared by translating the Spanish side of the Chinese-Spanish *United*

[5]Jieba is a popular open source Chinese segmentation library: https://github.com/fxsjy/jieba.
[6]`tensor2tensor` source code is available at https://github.com/tensorflow/tensor2tensor/. For this work, version 1.2.9 was used.

Nations (UN) corpus into Catalan and *pseudo-corpus2* is prepared by translating the Spanish side of the Spanish-Catalan *El Periódico* corpus into Chinese.

Table 4. Translation results (uncased BLEU scores). In bold, best results.

Language	System	Train.data	Test data	BLEU
Zh→Es	Direct	UN	UN	46.25
Es→Ca	Direct	EP	EP	87.04
Zh→Ca	Cascade	UN	UN	38.58
Zh→Ca	**Pseudo-corpus1**	UN	UN	**38.92**
Es→Zh	Direct	UN	UN	44.16
Zh→Ca	Pseudo-corpus2	EP	UN	22.45

There are three main conclusions that we can extract from what is shown in Table 4. First, the *pseudo-corpus1* approach outperforms the *pseudo-corpus2* approach despite the fact that the target side of the former is synthetic. This may be due to the high quality of the Spanish-to-Catalan system which is used to generate the target side of *pseudo-corpus1*. Note that results from Spanish-to-Catalan with our Transformer system outperform state-of-the-art previous results [8]. Additionally, the test set belongs to the *United Nations* domain, which is different than the domain of the training data of *pseudo-corpus2*. This way, when choosing among the two pseudo-corpus versions, the final intended domain for the system has to be taken into account, together with the relative quality of the translation system used to generate the synthetic side of the pseudo-corpus.

Second, the pseudo-corpus approach performs slightly better than the cascade approach, coherently to the case of statistical machine translation [13].

Third, we are showing that Chinese-to-Catalan obtains a 38.92 BLEU which is remarkable specially taking into account that translation is performed without any direct parallel data.

Table 5 shows some examples comparing best Catalan translation (using best pseudo-corpus pivot approach) together with the Spanish translation. Translation quality has reasonable adequacy and fluency in both cases. Note that only extra generation of pronouns is shown in the second example.

7 CONCLUSIONS

This paper brings together the challenge of addressing Chinese-to-Catalan translation with the latest neural machine translation techniques contrasting a couple of pivot approaches (cascade and pseudo-corpus). Results show that the pseudo-corpus approach outperforms cascade and we reach a high 38.92 BLEU in the Chinese-to-Catalan task.

For our experiments, it was necessary the development of a gold standard for Chinese-Catalan. Given that the text is from the *United Nations*, this gold standard can also be extended for the other *United Nations* official languages (Arabic, English, French, Russian and Spanish). This gold standard from the 5-official *United Nations* official languages to Catalan is made freely available for the community.

ACKNOWLEDGMENTS

This work is supported by the Spanish Ministerio de Economía y Competitividad and European Regional Development Fund, through the postdoctoral senior grant *Ramón y Cajal*, the contract TEC2015-69266-P (MINECO/FEDER, UE) and the contract PCIN-2017-079 (AEI/MINECO). This work is also supported in part by an Industrial PhD Grant from the Catalan Agency for Management of University and Research Grants (AGAUR) and by the United Language Group (ULG).

Table 5. Sample end-to-end Chinese-to-Catalan and Chinese-to-Spanish translations.

Chinese	此外，还必须制定适当的政策整合指标，同时考虑到 经济合作与发展组织在这一领域开展的工作。
Catalan	A més a més , s'han d'elaborar indicadors adequats d'integració de polítiques , tenint en compte la feina de l'Organització de Cooperació i Desenvolupament Econòmics en aquesta esfera.
Spanish	Además, es necesario elaborar indicadores adecuados de integración de las políticas, teniendo en cuenta la labor de la Organización de Cooperación y Desarrollo Económicos en este ámbito.
Chinese	它也决心遏止恐怖主义的危险和恐怖，以期捍卫叙利亚公民及其荣誉， 并还击对我国及人民能力的攻击。
Catalan	També està decidida a posar fi al perill i el terror del terrorisme , amb vista a defensar els ciutadans sirians i el seu honor i a respondre als atacs contra e l **nostre*** país i les capacitats del seu poble.
Spanish	También está decidida a luchar contra los peligros y el terror del terrorismo, a fin de defender a los ciudadanos sirios y a su honor y responder a los ataques contra la capacidad de **mi*** país y su pueblo.
Chinese	对人道主义需求十分巨大且与日俱增
Catalan	Les necessitats humanitàries són enormes i creixents.
Spanish	Las necesidades humanitarias son enormes y aumentan día a día.

REFERENCES

[1] Maite Ardèvol. 2006. Informe Anual OME 2006: Tendències de futur i noves realitats. (2006).

[2] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised Machine Translation Using Monolingual Corpora. *CoRR* abs/1711.00041 (2017).

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[4] Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631* (2016).

[5] Ibana Casaburi. 2016. Chinese International Investment. (2016).

[6] Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint Training for Pivot-based Neural Machine Translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 3974–3980.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[8] Marta R. Costa-jussà. 2017. Why Catalan-Spanish Neural Machine Translation? Analysis, comparison and combination with standard Rule and Phrase-based technologies. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics, Valencia, Spain, 55–62.

[9] Marta R. Costa-Jussà, David Aldón, and José A. Fonollosa. 2017. Chinese—Spanish Neural Machine Translation Enhanced with Character and Word Bitmap Fonts. *Machine Translation* 31, 1-2 (June 2017), 35–47.

[10] Marta R. Costa-Jussà and Jordi Centelles. 2015. Description of the Chinese-to-Spanish Rule-Based Machine Translation System Developed Using a Hybrid Combination of Human Annotation and Statistical Techniques. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 15, 1, Article 1 (Nov. 2015), 13 pages.

[11] Marta R. Costa-jussà and C. Escolano. 2016. Morphology Generation for Statistical Machine Translation using Deep Learning Techniques. *CORR, arXiv:1610.02209* (2016).

[12] Marta R. Costa-Jussà, José A. R. Fonollosa, José B. Mariño, Marc Poch, and Mireia Farrús. 2014. A Large Spanish-Catalan Parallel Corpus Release for Machine Translation. *Computing and Informatics* 33, 4 (2014), 907–920.

[13] Marta R. Costa-jussà, Carlos A. Henríquez Q, and Rafael E. Banchs. 2012. Evaluating Indirect Strategies for Chinese-Spanish Statistical Machine Translation. *J. Artif. Int. Res.* 45, 1 (Sept. 2012), 761–780.

[14] Marta R. Costa-jussà, Noé Casas, and Maite Melero. 2018. English-Catalan Neural Machine Translation in the Biomedical Domain through the cascade approach. In *Proceedings of the 11th Language Resources and Evaluation Conference of the European Language Resources Association.*

[15] John DeFrancis. 1984. The Chinese Language: Fact and Fantasy. (1984).

[16] Mireia Farrús, Marta R. Costa-Jussà, José B. Mariño, Marc Poch, Adolfo Hernández, Carlos Henríquez, and José A. Fonollosa. 2011. Overcoming Statistical Machine Translation Limitations: Error Analysis and Proposed Solutions for the Catalan—Spanish Language Pair. *Lang. Resour. Eval.* 45, 2 (May 2011), 181–208.

[17] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122* (2017).

[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[19] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *CoRR* abs/1611.04558 (2016). arXiv:1611.04558

[20] Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation.* Association for Computational Linguistics, Vancouver, 28–39.

[21] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised Machine Translation Using Monolingual Corpora Only. *CoRR* abs/1711.00043 (2017).

[22] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[23] Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *International Conference on language Resources and Evaluation.*

[24] Gema Ramírez Sánchez, Felipe Sánchez-Martínez, Sergio Ortiz Rojas, Juan Antonio Pérez-Ortiz, and Mikel L. Forcada. 2006-11. Opentrad Apertium open-source machine translation system: an opportunity for business and research. (2006-11).

[25] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Berlin, Germany, 1715–1725.

[26] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems.* 3104–3112.

[27] Antonio Toral and Víctor M Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *arXiv preprint arXiv:1701.02901* (2017).

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 6000–6010.

[29] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (23-28), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Paris, France.

Received February 2007; revised March 2009; accepted June 2009